Data Abstraction Framework for Heterogeneous Logs of Scientific Clusters

Project Description

Large science projects are increasingly relying on thousands of CPUs to produce and analyze petabytes of data. Their executions usually involve thousands of concurrent operations of data accesses, data movement, and computation. Due to the lack of common framework that accommodates different types of measurement logs, the analyses on these data usually involve one type of log separated from other types of logs. The usage analysis from the logs are hard to find in HPC applications and scientific clusters. Understanding the characteristics of the complex application workflows and analyze their executions are challenging for various reasons. The concurrent data accesses may compete with each other and with other jobs for accessing shared data storage and networking resources. The storage and memory hierarchies on the current generation of hardware are very complex, and therefore have performance or reliability characteristics that are sometimes unexpected. As the executions involve multiple resources in various layers in the system, the analysis requires inspecting all different types of logs that are involved in the executions. Therefore, it is difficult to analyze them without the help from the common data abstraction of these different types of logs from a user-friendly framework support for the data abstraction.

This project is motivated by the observations that different types of logs from scientific clusters are available but they are not frequently used due to lack of a common data abstraction. We aim to develop a framework to provide the common data abstraction for different types of measurement data. It can support multiple analysis use cases such as joining different types of logs and discover meaningful insights about execution fluctuations and unexpected behaviors.

Task Goals

The main research goal is to develop a framework to provide a common data abstraction for different types of logs from multiple scientific clusters such as NERSC or SLAC. It includes following items:

- Development of a framework to load different types of logs using open-source Elastic search module
- Development of a common data abstraction to provide accessible analysis on different types of logs

Task Requirements

- Proficient in a programming language, such as python.
- Good problem solving skills and communication skills
- Enthusiastic about tackling research challenges and solving problems

About the group

The Scientific Data Management (SDM) group at Lawrence Berkeley National Laboratory develops technologies and tools for efficient data access, data storage, data analysis, and management of massive scientific data sets. We are currently developing storage resource management tools, data querying technologies, in situ feature extraction algorithms, data analysis algorithms, along with software platforms for exascale data. The group also works closely with application scientists to address their data processing challenges. These tools and application development activities are backed by active research efforts on novel algorithms for emerging hardware platforms.